Brief Papers

NCMOS: A High Performance CMOS Logic

Rajendra Kumar

***Abstract*—CMOS has been the mainstay technology for VLSI design for the last several years. However, recently, BiCMOS technology has been proposed for speed critical applications. In this paper we propose a new circuit structure called NCMOS, which employs a low *Vt* NMOS transistor in place of the bipolar transistor, and provides significantly higher speed than a conventional CMOS design. This is realized at the cost of only one extra masking step, compared to 4–5 extra masks for a full BiCMOS process.**

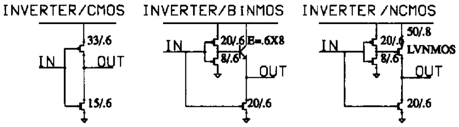

Fig. 1. Inverters in (a) CMOS, (b) BiNMOS, and (c) NCMOS.

I. Introduction

For VLSI design, CMOS technology is a natural choice because of zero DC power dissipation. However, because MOS devices in general, and PMOS devices in particular, do not have as high a current drive as an npn bipolar transistor, BiCMOS circuits have been used recently [1]. The BiCMOS circuits preserve the low power dissipation characteristic of the CMOS, and also provide a higher current drive of the bipolar transistor. But this is realized at the cost of significantly increased process complexity, and cost. In addition, there is the problem of non-rail-to-rail swing at the output of a gate because of the bipolar diode drops. This is especially significant for a full BiCMOS gate, where the diode drop appears both when the output goes high and low. This problem can be somewhat alleviated by use of BiNMOS gates, Fig. 1(b), where the pulldown device has been replaced by an NMOS transistor. In this case, the output low transition would go all the way to 0 V. Circuit solutions have been proposed, e.g. [3], which enable the output of a BiCMOS gate to go rail-to-rail. But all these techniques effectively bypass the bipolar transistor, either by an MOS transistor, or a resistor to pull the output rail-to-rail. Even though these techniques usually work in the 3–5 V range, it is not clear whether these techniques would be effective down to 1.5 V.

The structure of BiNMOS gate is interesting to analyze. Fig. 1(a)–(b) shows the comparison of a CMOS versus a BiNMOS inverter. The pulldown structure of the two gates is quite similar in the sense that both gates use NMOS devices as pulldowns. But the pullup structure is quite different. The CMOS case uses PMOS devices as pullup, whereas the BiNMOS case uses an npn transistor. In the CMOS case, the PMOS devices implement the logic function (inverter in this case), as well as drive the output load. But in the BiNMOS gate, the logic function is implemented by a CMOS inverter whose output drives the base of the npn transistor. The npn transistor itself drives the output load. In any case, the net result of moving from the CMOS gate to the BiNMOS gate is that the slower PMOS device has been replaced by a faster npn transistor! But it turns out that logically speaking, a regular NMOS device should work as well in place of the npn transistor. The NMOS device in general has a 2 X to 3 X mobility advantage compared to a PMOS device. So a regular NMOS device in place of the pullup npn transistor should provide a higher speed than a pure CMOS gate. The next section will show such comparisons. But there is one problem with the use of the regular NMOS device as a pullup. This has to do with the *Vt* drop, and the body effect of the pullup NMOS. Because of these effects, the output high could be as much as 1–1.5 V below *VDD*. This problem becomes particularly acute when the *VDD* supply is lowered to 3.3 V for submicron processes. However, if we were to use a low threshold voltage ($Vt \sim 0.2$ V) NMOS device (called LVNMOS) as a pullup, the *Vt* drop including the body effect is small (~ 0.4 V), and the speed of the gate is still significantly faster than a pure CMOS gate. We call such circuits NCMOS, Fig. 1(c), because these circuits use CMOS structures which are augmented by a low *Vt* NMOS device as the final pullup. We should point out that 0 *Vt*, or depletion devices have been used in pure NMOS circuits like superbuffers in the past, [2]. But such circuits were "ratioed," and their output low level (Vol) was determined by a ratio of the on resistance of the pullup and pulldown and hence the low output did not go down to 0 V. But more significantly, these NMOS circuits had a DC current flow when the output was low, and hence a high DC power dissipation. The NCMOS circuits we propose in this paper are "ratioless," just like CMOS, and their output low voltage goes all the way to 0 V. Besides, there is no DC power dissipation. The only DC current flow is due to leakage. The leakage current would of course be dependent on the *Vt* of the low threshold device, and would be quite small compared to a 0 *Vt* or depletion mode device. Interestingly, the leakage current can be traded with the *Vt* drop of the pullup device

Manuscript received March 19, 1993; revised January 19, 1994.

The author is with Hewlett Packard Laboratories, Computer Systems Laboratory, Palo Alto, CA 94304.

IEEE Log Number 9401444.

0018-9200/94$04.00 © 1994 IEEE

in NCMOS circuits. We find that even for regular NMOS pullups (instead of the low Vt devices), the gate delay is still significantly faster than the pure CMOS case.

II. Results

Fig. 1(a)–(c) shows the schematic details of three classes of circuits: CMOS, BiNMOS and NCMOS. All three circuits are so configured that their input capacitance is the same. Table I shows the SPICE results of gate delay for various loads. The device models for the regular NMOS, PMOS and bipolar devices were obtained from measured data, and the low Vt NMOS device models were extracted from a PISCES simulation. The low Vt device (called LVNMOS) had a deep threshold implant (common to the regular Vt device) to suppress subthreshold leakage, as well as raise the threshold voltage to 0.2 V. It can be seen from Table I that the NCMOS gate has a significantly higher speed (~25–43%) than a CMOS gate, and the BiNMOS gate provides an additional 8–13% speedup over NCMOS, for different loads. The channel length of the LVNMOS device was kept somewhat larger (0.8 μm) than the minimum channel length of 0.6 μm for the process, in order to keep the leakage down. f_t of the npn bipolar transistors used in the BiNMOS gates was 15 GHz, and the h_{fe} was 100. The nominal Vt for the regular PMOS and NMOS devices was −0.65 V, and 0.65 V, respectively. The nominal subthreshold leakage current of the low Vt device was 0.1 μA for a 10 μm/0.8 μm (W/L) device. The worst case leakage (fast models, high VDD of 4.2 V, high temperature of 85°C) was 1.56 μA for a 10 μm/0.8 μm device. Therefore for a pullup device of 50 μm/0.8 μm, the worst case subthreshold leakage would be 7.8 μA. This would amount to a dc power dissipation of 7.8 μA × 4.2 V = 32.8 μW per gate. The ac power dissipation for such a gate with 0.3 pf load, and a 100 MHz operation would be $= 0.3\ \text{pf} \times (4.2 - 0.4) \times 4.2 \times 100$ MHz = 478.8 μW. As a result, the dc power due to subthreshold leakage of the low Vt device is about 6.85% of the ac power dissipation. Even though the absolute dc power dissipation of an NCMOS gate would be higher than a comparable CMOS gate, the total power dissipation (including ac power) would be only marginally more for the NCMOS case. Assuming a hypothetical chip with 100 K gates, with 10% of the gates being NCMOS, the total worst case dc leakage current (maximum Vdd, maximum temperature, fast process) for the whole chip would be about $10,000 \times 7.8$ μA = 78 mA. This might appear to be a significantly higher leakage than a fully static, pure CMOS chip, but again the impact on total power dissipation would be small for most cases. One exception may be the battery operated systems, which go to sleep mode when idling, where the extra dc leakage may affect the battery life between charges. We have no statistical data on the Vt spread across process corners for the low Vt device. But we found that the Vt spread for our regular Vt devices is 0.65 V +−60 mV. Also, the Vt spread for natural Vt devices (no channel implant was −0.15 V +50 mV. Based on this data, it would appear that the Vt spread for the low Vt device should be comparable to a regular Vt device. But, in the author's view, more data needs to be collected on the Vt spread of low Vt devices, in order to get quantitative information. Nonetheless, we would like to point out that useful chips have been built using low Vt devices, e.g., [4].

TABLE I
Inverter Delays for BiNMOS, NCMOS, and CMOS. (Worst Case Delays: Vdd=3.6 V, Temperature=85C, Slow Models). (*)–Speedup Over CMOS=[td avg(CMOS)–td avg]/td avg(CMOS)

CL, pf	Circuit	td h, ps	td l, ps	td avg, ps	% Speedup over CMOS (*)
0.3	BiNMOS	184	94	139	38.5
	NCMOS	220	118	169	25.2
	CMOS	228	224	226	0
0.5	BiNMOS	217	142	179.5	45.4
	NCMOS	258	174	216	34.4
	CMOS	335	323	329	0
1.0	BiNMOS	289	270	279.5	51.9
	NCMOS	350	309	329.5	43.3
	CMOS	602	560	581	0

TABLE II
NAND2 Delays for: BiNMOS, NCMOS, and CMOS. (Worst Case Delays: Vdd=3.66 V, Temperature=85C, Slow Models). (*)–Speedup Over CMOS=[td avg(CMOS)-td avg]/td avg (CMOS). Note: NCMOS-I Gates Use LVNMOS Devices for Only the Output Pullup Function, Whereas NCMOS-II Gates Use LVNMOS Devices for Both Output Pullup and Pulldown Functions.

CL, pf	Circuit	td h, ps	td l, ps	td avg, ps	% Speedup over CMOS (*)
0.3	BiNMOS	246	88	167	24.4
	NCMOS-I	288	105	196.5	11.1
	NCMOS-II	288	64	176	20.4
	CMOS	249	193	221	0
0.5	BiNMOS	275	121	198	34.1
	NCMOS-I	333	152	242.5	19.3
	NCMOS-II	333	86	209.5	30.3
	CMOS	338	263	300.5	0
1.0	BiNMOS	341	219	280	43
	NCMOS-I	417	260	338.5	31.1
	NCMOS-II	418	144	281	42.8
	CMOS	557	426	491.5	0

For an inverter, the ratio of areas for CMOS versus BiNMOS versus NCMOS is .79:1:1.54, and for a NAND2 the ratio is .92:1:1.21. In most cases, the impact of this area increase for the NCMOS gate should be minimal, because one would tend to use NCMOS gates for relatively high load capacitances. Besides, in processor type of chips area is governed more by size of on-chip caches, busses, I/O etc.

Fig. 2 shows the circuits for a NAND2 for the three cases. Table II shows SPICE simulation results. Again it can be seen that the NCMOS-I gate (LVNMOS used as a pullup only) is significantly faster than a pure CMOS case, and comes close in performance to a BiNMOS gate.

The performance of the NCMOS circuits can be further enhanced by using the low Vt device even for the pulldown (NCMOS-II gates). Use of a low Vt device for pulldown provides higher gate drive, and should speed up the high to low transition of the output. In order to keep the dc leakage small, the IR drop on the power busses should be kept small. This would prevent the spurious turn-on of the low Vt devices. Table II shows the gate delays for a NAND2 for the three

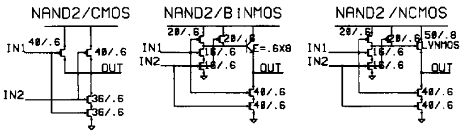

Fig. 2. NAND2 schematics in (a) CMOS, (b) BiNMOS, and (c) NCMOS.

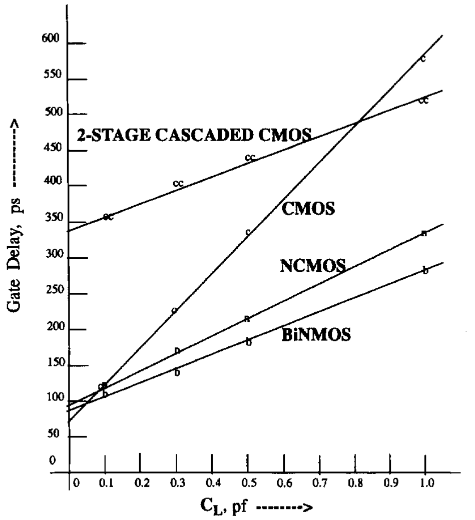

Fig. 3. Speed dependence of CMOS, BiNMOS, NCMOS inverters, and a two-stage cascaded CMOS inverter chain as a function of fan-out.

cases. In this case, the NCMOS circuit speed (NCMOS-II gates) is within 4% of the BiNMOS circuit.

Fig. 3 shows the speed dependence of CMOS, BiNMOS, NCMOS (LVNMOS pullups only)-inverters, and a two-stage cascaded CMOS inverter chain as a function of fan-out.

We also compared the period of oscillation of a three-stage ring oscillator, with the same input, and output capacitances for the CMOS, NCMOS (LVNMOS pullups only), and BiNMOS circuits. Again, the NCMOS scheme is significantly faster (37.7%) than CMOS. The BiNMOS scheme is 47.9% faster than CMOS.

In order to design circuits using NCMOS, it would be helpful to have an understanding of some of the design tradeoffs. The high output level of an NCMOS gate would be equal to $VDD - Vt$(LVNMOS). Since this output could drive the PMOS device of a regular CMOS gate, it would be desirable to keep the Vt(LVNMOS) smaller than the absolute value of the Vt of the PMOS device, for all cases. This would ensure that the leakage current of the PMOS device does not become large. Alternatively, one could adapt the NCMOS circuit to make it rail-to-rail, by adding a PMOS pullup device in parallel with the low Vt NMOS pullup, as is commonly done in BiCMOS circuits. The NCMOS logic is an inverting logic, i.e., the common gate types realized would be inverters, nand gates, nor gates etc. Such circuits when cascaded maintain an output high level of $VDD - Vt$(LVNMOS), irrespective of the number of stages. This is because, the gate voltage of the LVNMOS pullup is always rail-to-rail, as it is driven by the output of a pure CMOS gate, Figs. 1 and 2. The cascading of the NCMOS gates still provides a significant speed advantage over CMOS, as shown by the ring oscillator data above. Also, cascading does not degrade noise margin, because the output swing is always between 0 V and $VDD - Vt$ (LVNMOS), irrespective of the number of cascaded stages. In terms of process complexity, the NCMOS gate would require one extra mask compared to CMOS, but significantly fewer masks than BiCMOS. As a result, the theoretical yield on an NCMOS process would be somewhat lower than of a comparable CMOS process. We do not believe that the NCMOS yield would be significantly lower, but on the other hand, the performance level would be significantly higher compared to CMOS. Theoretically, there is no hard lower limit on the threshold voltage of the low Vt device. But in practical terms, we think, a lower Vt limit of 0.1 V should keep leakage currents to low enough levels to make these devices useful in VLSI structures. In terms of scalability, we think, these circuits should scale well with supply voltages at least up to 1–1.5 V. Again, this may be a practical consideration, because unlike the BiCMOS circuits where the diode drop does not scale, the Vt of an NMOS device can scale all the way to 0 V, or even negative values. Another point to note in designing NCMOS circuits is the spread in the Vt values for the LVNMOS device. As a result worst case simulations across process, voltage, and temperature should be done (as in pure CMOS designs) to ensure that the NCMOS circuits work under all operating conditions.

III. Conclusion

A new high performance circuit (NCMOS) has been proposed which has significantly faster speed than CMOS. This circuit uses only MOS devices, is scalable with voltage, and uses only one extra mask compared to a regular CMOS process. This circuit overcomes some of the important shortcomings of BiCMOS, and yet provides comparable speed.

Acknowledgment

The author would like to acknowledge the help of Paul Vande Voorde for providing the PISCES data for the low Vt device and Peter George for providing the device models for this device. The author would like to thank Gene Emerson for making several suggestions regarding this work. He would also like to express his gratitude to Bill Worley and Dick Lampman for their continued encouragement and support. Finally he wishes to thank the reviewers for their helpful comments.

References

[1] A. Alvarez, *BiCMOS Technology and Applications*, Norwell, MA: Kluwer, 1989.

[2] L. A. Glasser and D. W. Dobberpuhl, *The Design and Analysis of VLSI Circuits*, Reading, MA: Addison-Wesley, 1985, pp. 26–28.

[3] H. Hara et al., "0.5 μm 2 M-Transistor BipnMOS Channelless Gate Array," in *ISSCC Dig. Tech Papers*, 1991, pp. 148–149.

[4] G. Kitsukawa et al., "256 Mb DRAM technologies for file applications," in *ISSCC Dig. Tech Papers*, 1993, pp. 48–49.